Development of a Modern Greek Broadcast-News Corpus and Speech Recognition System

Jürgen Riedler
SAIL Labs Vienna, Austria
juergen@sail-labs.at

Sergios Katsikas
Department of German Linguistics
University of Pécs, Hungary
Katsikas@btk.pte.hu

Abstract

We report on the creation of a Modern Greek broadcast-news corpus as a pre-requisite to build a large-vocabulary continuous-speech recognition system. We discuss lexical modelling with respect to pronuciation generation and examine the effects of the lexicon size on word accuracies. Peculiarities of Modern Greek as a highly inflectional language and their challenges for speech recognition are discussed.

1 Introduction

Modern Greek *Koine* or Standard Modern Greek, the official language of Greece and Cyprus, is the latest variety of Europe's oldest literary language following Mycenian, Ancient, Hellenistic, and Byzantine Greek. Research objectives within the REVEAL THIS[1] project comprise also the development of a Modern Greek (MG) automatic speech recognition (ASR) system. In constrast to recent efforts on MG ASR focussing on dictation (Digalakis et al., 2003), our interests are in the broadcast news domain.

After providing a short linguistic overview of MG we specify the prerequisites for ASR, which would be: audio recordings with corresponding transcriptions to train acoustic models, text corpora for language modelling, and recognition lexicon inclusive pronunciation generation. Finally we disclose word error rates of experiments employing various recognition dictionaries and discuss major problems of lexical and language modelling for a highly inflectional language.

[1] **Re**trieval of **V**id**e**o **A**nd **L**anguage for **T**he **H**ome user in an **I**nformation **S**ociety – funded by the IST Frame Programm 6/2003/IST/2.*Scientific and technological objectives:* 1) Augmentation of the content of multimedia documents with entity, topic, speaker, and fact information; 2) Development of cross-media and cross-language representations; 3) High-level functionalities, like search, retrieval, categorization, and summarization, from 1) and 2).

2 Notes on Modern Greek structure

In the following we briefly present a linguistic introduction into MG - see (Katsikas, 1997), (Mackridge, 1985) and references therein - and comment on its implications to ASR.

2.1 Phonological system

The phonological system of MG consists of five vowel phonemes: /a/, /ɛ/, /i/, /o/, /u/ and 20 consonant phonemes: the plosives /p/, /b/, /t/, /d/, /k/, /g/, the fricatives /f/, /v/, /θ/, /ð/, /s/, /z/, /x/, /ɣ/, the affricates /t͡s/, /d͡z/, the nasals /m/, /n/, the lateral /l/ and the apical trill /r/. The most important allophone-generating phonological processes are:

- palatalisation of /k/, /g/, /x/, /ɣ/ to [c], [ɟ], [ç], [j] before /i/ or /ɛ/
- /k/, /g/, /x/, /ɣ/, /n/, /l/ merge with following glide [j] (non-syllabic allophone of /i/) to palatals: [c], [ɟ], [ç], [j], [ɲ], [ʎ], *e.g.* εννιά /ɛniˈa/ → *[ɛˈnja] → [ɛˈɲa]
- sonorisation of /p/, /t/, /k/, /t͡s/ to [b], [d], [ɡ], [d͡z] after /n/, often with denasalisation in informal speech, *e.g.* τον πατέρα /ton paˈtɛra/ → [tombaˈtɛra] or [tobaˈtɛra]
- regressive assimilation of place of articulation of /n/ to the following consonant
  - /n/ → [m] before /p/, /b/, see former example
  - /n/ → [ŋ] before /k/, /g/, /x/, /ɣ/, *e.g.* τον Κώστα /ton ˈkosta/ → [toŋˈgosta] or [toˈgosta]
- sonorisation of /s/ to [z] before voiced consonants, *e.g.* της λέω /tis ˈlɛo/ → [tizˈlɛo]

Within syntactic phrases (*e.g.* article - noun - possesive pronoun) certain phonological processes usually extend even across word boundaries (see examples above), but only if there is no pause between the words.

This can cause homophony of phrases, *e.g.* [timˈbira] or [tiˈbira] could mean both την μπίρα "the beer {acc.}" or την πήρα "I picked her up/I called her *etc.*", and represents an almost inevitable source of word errors for ASR (*cf.* Section 4).

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 380–383

2.2 Prosody

The functional load of prosodic features in MG is extremely high, since word stress and intonation are highly distinctive. There are hundreds of prime-stress minimal pairs (*e.g.* πότε "when" *vs.* ποτέ "never"), stress fulfills various morphological functions and moreover, intonation patterns provide in most cases the only distinction between declarative clauses and yes-no questions (*e.g.* [o ˈjanis ˈin(ɛ) ɛˈðo↘] "John is here" *vs.* [o ˈjanis ˈin(ɛ) ɛˈðo↗] "Is John here?").

This is the reason why we introduced word stress as a part of suprasegmental structure into our phone sets, see Section 3.

2.3 Morphology

MG is a prototypical inflectional language, *i.e.* a potentially huge number of different word forms may be derived from one basic stem (lemma). In particular verb inflection is very rich: by combining two stems, three sets of endings, a few modal particles, an auxiliary verb and the participle, every active verb can produce about 200 forms, if we take all syntactically defined categories (three aspects, six moods, eight tenses, *etc.*) into account, despite of (partial) homonymies. This number is twice as big for verbs that exhibit a medio-passive voice, which is formed synthetically. Different verb forms can differ from each other in the ending, in accentuation as well as in the stem (there are also irregular verbs with suppletive roots, *e.g.* βλέπω [ˈvlɛpo] "I see" *vs.* είδα [ˈiða] "I saw"), and finally, active verbs consisting of two syllables have in past tense a sort of prefix (augment) carrying the stress on the antepenultimate syllable. Nouns show, depending on their inflectional class, between 4 and 7 different forms, adjectives about 40 (including comparative and elative). Due to ambiguities of various morphological rules and the bistructurality of MG (parallel use of old and new forms), inflectional forms are often hardly predictable.

MG word formation processes are very complex though not very productive. Various mutations of morphemes and bistructurality prevent the predictability of derivatives and compounds. For example, the stems within the verb forms κλέβω "I steel", έκλεψα "I stole" do not evidently imply those in derivatives like κλέφτης "thief", κλοπή "theft" or in a compound like κλεπτομανής "cleptomaniac".

Since syntactic relations between constituents of a sentence are mostly expressed by inflection, MG constituent order is fairly free Word order has rather a pragmatic than syntactic function (e.g. topicalisation).

It is obvious from the above that inflection as well as syntactical freedom present outstanding demands on lexical and language modelling.

3 Phonetic transcription and lexicon

MG grapho-phonemic correspondences are mostly unambiguous from grapheme to phoneme, *i.e.* the pronunciation of written text is predictable to a high degree. However, as a result of historical spelling some phonemes comply with more than one grapheme (*e.g.* /i/ may be represented by six different graphemes: ⟨ι⟩, ⟨η⟩, ⟨υ⟩, ⟨ει⟩, ⟨οι⟩, ⟨υι⟩), hence the text-to-speech task or pronuciation generation, respectively is less problematic than ASR.

Recognition dictionaries map lexical words to their corresponding phonetical transcriptions. This was accomplished by an automatic grapheme-to-phoneme (g2p) conversion applying about 70 rules in consideration of:

- structure words like τον, την (male and female definite article in accusative) and very frequent monosyllabic words were transcribed manually because of their manifold phonetical realisations
- the ⟨γγ⟩-digraph resulting from 'learned' formations of the prefixes {εν-, συν-} and stems with initial /ɣ/, *e.g.* έγγραφο "document", is phonetically transcribed as [ŋɣ] (in contrast to [ŋg] as usual)
- company or product names and acronyms written in latin characters (*e.g.* BBC, Unesco, Löwenbräu) also had to be transcribed manually

g2p makes use of the following phone inventory: 26 consonants (except for affricates like /t͡s/, /d͡z/, which were separated as /t/+/s/, /d/+/z/, respectively), 5 vowels, the non-syllabic /i/ plus one additional phone for every stressed vowel and 4 artificial phones (SILence, BReaTh, LIPsmack, GaRBage).

Aside from the phonological processes described above (*cf.* Section 2.1) the following phenomena were found to be relevant for phonetic transcription:

- pronunciations of the consonantal digraphs ⟨μπ⟩, ⟨ντ⟩, ⟨γκ/γγ⟩ within words vary between [b], [d], [g] and [mb], [nd], [ŋg] (not at word beginnings) due to regional, stylistic, and individual differences

- digraphs ⟨αυ⟩, ⟨ευ⟩, ⟨ηυ⟩ are pronounced as [af], [ef], [if] before voiceless consonants and as [av], [ev], [iv] before voiced consonants or vowels
- within pronunciations of the digraphs ⟨αύ⟩, ⟨εύ⟩, ⟨ηύ⟩, the vowel has to be stressed, although for reasons of orthography the written accent is put on the consonantal component

Obeying the specified phonological rules lead to 1.9 pronunciations per lexeme on average.

4 Experimental setup

4.1 Corpora

Experiments were carried out using audio recordings (mono, 16kHz sampling rate, 16 bit resolution) of various news shows broadcasted via the Greek satellite-TV channel EPT . Transcription into text as well as XML-annotation (timing, speaker turns and names, topics, non-speech utterances, *etc.*) of the collected audio data was done at ILSP[2]. The recorded data comprise

- $\sim$ 27000 pure speech segments (utterances)
- $\sim$ 1200 individual speakers of which $\sim$ 300 could be identified by name
- $\sim$ 1500 segments (stories) annotated according to a topic hierarchy derived from Reuters

and were randomly divided into a training set of 36^h05^{min} and a disjoint test set of 1^h35^{min}.

Two corpora made up of newspaper texts of approximately 25 million words altogether were provided by ILSP and had to undergo several preprocessing steps in order to obtain clean and convenient text for language modelling. This gave an exhaustive word list of about 350k different lexical terms of which 200k occur more than once, see *e.g.* (Oikonomidis and Digalakis, 2003) for a comparison with other European languages.

4.2 Recognizer

Acoustic models are context-dependent triphone (1984 codebooks) and quinphone models (76432 codebooks) derived from mel-frequency cepstra (cepstral coefficients up to 14^{th} order as well as their first and second derivatives) extracted from the audio. Several normalisation and adaptation techniques like cepstral mean subtraction are applied on a per-utterance base. The phone models are continuous-density Hidden Markov Models with state-tied Gaussian mixtures employed in two subsequent decoder passes.

[2]Institute for Language and Speech Processing ⟨http://www.ilsp.gr⟩

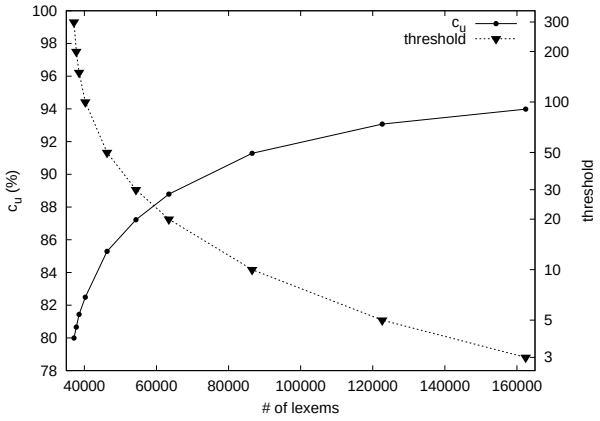

Figure 1: *Lexical coverages c_u obtained by using all words from the audio transcripts supplemented by those words with occurences of more than a minimum-threshold in the text corpus.*

We adopted back-off trigram language models with modified Witten-Bell smoothing. Language models were trained on the audio transcripts as well as the newspaper corpora, in which the audio data were given a higher weight (because audio vocabulary and n-gram inventory is supposed to be more similar to the ASR's actual operational area).

The decoder is part of the next-generation SAIL Labs Media Mining System. It is designed to run in real-time on state-of-the-art PC hardware (details will be published elsewhere).

4.3 Experiments

The recognition lexicon was assembled by taking all words from the audio transcripts as a basis, and extending it by those words of the text corpora with frequencies higher than a given threshold. Figure 1 depicts lexical coverages on the test set as a function of the number of lexical terms. In addition one can read off that coverages due to a cut-off of no more than 3 yields a dictionary of about 160k lexems, *i.e.* only the inclusion of words with rather small unigram probability, lead to coverages generally reported for recognition dictionaries of comparable utility, *cf.* (Oikonomidis and Digalakis, 2003).

We tested several ASR systems with respect to lexicon size and got almost constant word error rates of about **38%** for recognition lexicons with **90k-160k** entries, corresponding to lexical coverages of greater than **90%**, see Figure 2. Additional words of low frequency don't reduce word error

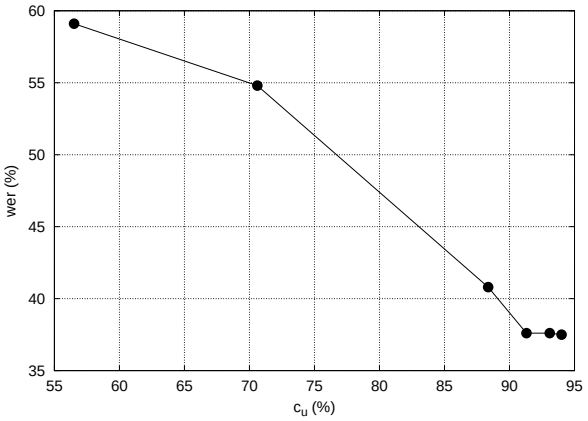

Figure 2: *Word error rates versus lexical coverage of the recognition dictionary on the test set.*

rates further as support by the language model collapses due to missing trigrams. This is also reflected in trigram perplexity figures ranging within **320-330**.

Apart of problems due to out-of-vocabulary words, the most frequent types of errors are insertions and deletions of common, poorly articulated, short words like negative and modal particles, articles, prepositions, and conjunctions. Another source of error is provoked by homophonies of word transitions within different word sequences, which cause wrong word boundary settings, *e.g.* note the displacement of initial [s] in the REFerence (Σ, σ) to final [s] in the HYPothesis (ς):

REF: ...στη Λεωφόρο Σπάτων στη ...
HYP: ...στη Λεωφόρος πάντως τη ...

A well endowed language model seems to be the only way out in this case.

5 Conclusions and Perspectives

Governed by the Modern Greek (MG) phonolocical and prosodical system we presented a grapheme-to-phoneme conversion for pronunciation generation necessary for ASR dictionaries. Several experiments were carried out employing language models and lexica of different extent. The resulting word error rate of around 38% may seem rather high, but is indeed within the ballpark for systems of comparable resources of training data. On the other hand, high perplexity values (compared to other European languages) is another indication of a rather difficult test set.

Concurrent ASR systems for inflectional languages, *e.g.* for Czech (Byrne et al., 2001), try to solve the problem of enormous vocabulary growth by performing automatic stemming and sophisticated morpheme-based language modelling. These techniques require grammatically tagged corpora and a morphological lexicon. However, as argued in Section 2, morphological decomposition is extremely non-systematic for Modern Greek and thus difficult to implement by means of rule-based stemming.

In (Oikonomidis and Digalakis, 2003) a maximum entropy language model incorporating n-gram (with Kneser-Ney smoothing) as well as stem constraints (word classification according to about 30k stems!) has been examined and a small but statistically significant improvement was achieved. Similar results were obtained from a factored language-modeling approach (Vergyri et al., 2004) with data-driven parameter optimization by genetic algorithms. Again small reductions of perplexity and word error rates are reported.

In view of the minor gain in performance using morphologically motivated language models, we expect considerable improvements by reducing the n-gram sparseness problem via incorporating much more language model data (keeping full form word lexica at the moment).

References

W. Byrne, J. Hajič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krbec and J. Psutka. 2001. *On Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language - Czech*. Proc. of Eurospeech 2001, Vol. 1: 487–490.

V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, V. Diakoloukas. 2003. *Large Vocabulary Continuous Speech Recognition in Greek: Corpus and an Automatic Dictation System*. Proc. of Eurospeech 2003: 1565–1568.

S. Katsikas. 1997. *Probleme der neugriechischen Graphematik aus der Perspektive des Fremdsprachenlernens* in H. Eichner et al. (eds): *Sprachnormung und Sprachplanung*: 419–474.

P. Mackridge. 1985. *Modern Greek Language - A Descritpive Analysis*. Oxford University Press, 1985.

D. Oikonomidis and V. Digalakis. 2003. *Stem-based Maximum Entropy Language Models for Inflectional Languages*. Proc. of Eurospeech 2003: 2285–2288.

D. Vergyri, K. Kirchhoff, K. Duh, A. Stolke. 2004. *Morphology-Based Language Modeling for Arabic Speech Recognition* Proc. of ICSLP 2004: 2245–2248